Bull. Nov. Comp. Center, Comp. Science, 34 (2012), 55–72

Methods of social network analysis

T. V. Batura

Abstract. This work represents a survey of the social network analysis problem. There are four main approaches: structural, resource-based, regulatory, and dynamic. To solve the problems in social network analysis, the following methods are used: graph and stochastic models, models of network evolution, methods involving ontologies, structural and relational models, machine learning methods, network visualization techniques, etc. The paper also describes the most popular computer social networks and some software applications to analyze them. It presents some possible paths of research: creation of an integrated theory of social networks, adaptation of methods of natural language text processing to the online content, etc.

Key words: Social network analysis, network model, graph of a network, data mining, centrality.

1. Introduction

Social networking as a social phenomenon emerged quite a long time ago. Social network analysis is used for studying the interactions between network members, predicting their behavior, their classification, modeling the flow of information in networks, etc. Currently, with the development of computer technologies, people have the opportunity to communicate virtually using social computer networks. Clearly, it is much faster and more convenient to analyze information and behavior of people in the process of communication in this format. Therefore, social network analysis raises great interest among researchers.

This paper outlines some possible directions of research and provides an overview of social networks: the main research areas are listed, some useful features of social networks are described, methods and algorithms used in various models of network analysis are outlined, the most popular to date social computer networks are described.

2. The main approaches to social network analysis

At present, there are four major approaches to social network analysis [1]: structural, resource-based, regulatory, and dynamic. Each of them solves quite a wide range of tasks and methods from different areas of knowledge.

In a **structural approach,** all members of a network are considered as vertices affecting the configuration of edges and other network participants. Attention is paid to the geometric form and intensity of interactions (weight

of the edges), so the following characteristics are investigated: relative positions of vertices, centrality, and transitivity of interactions. The structural analysis and analysis of the behavior of links use the techniques of statistical analysis, methods of community discovery, or classification algorithms. The behavior of nodes in a cluster and the behavior of typical temporal characteristics of social networks are studied, for example, changes in the network structure during its growth, or changes in the behavior and distribution of connected components of the graph.

Another issue examined in this section is community detection in social networks. The goal is to identify regions of a network with active engagement of participants. Algorithmically, this problem can be attributed to the problem of graph partitioning [2]. It is necessary to divide the network into dense regions based on the behavior of links between vertices. Social computer networks are dynamic, which causes difficulties in community detection. In some cases, it is possible to integrate the content network in the process of community discovery. Then the content helps us to identify groups of participants with similar interests.

The **resource-based approach** considers the ability of participants to attract individual and network resources to achieve certain goals and differentiates between parties in identical structural positions in a social network according to their resources. Individual resources can be knowledge, prestige, wealth, race, and gender. The network resources can be influence, status, scope and the nature of information. The main indicator of differences in the network resources is the strength of participant's structural position.

An important task in this area is the analysis of the social network content. The network content is a source for a wide range of applications aimed at data extraction and analysis. Using the network content, we can significantly improve the quality of conclusions in social network analysis, such as clustering and classification tasks. There are four types of content analysis [2].

1. The methods of random walks, for example, are used in the ***analysis of general information with arbitrary data types***. One of the best known algorithms that use similar methods is the PageRank algorithm for measuring the relative importance of Web documents: a page that is linked to many pages with high PageRank receives a high rank itself. If there are no links to a web page, there is no support for that page. Thus, PageRank is the method of calculating the page weight by counting the importance of references. This algorithm can also be used for search and classification of entities and participants in a social network and for assessing the probability of visiting any vertex. Naturally, the vertices which are better positioned from a structural point of view have a higher weight and, therefore, are more important.

Random walk methods may also be useful for merging participants into groups relative to the most influential participants.

2. The methods of data integration are used for ***analysis of sensors and stream mining*** in social networks. Many modern mobile phones support the capability to interact dynamically with each other in real-time depending on their location and status. They are used for obtaining information about a person or a combination of object properties that are tracked.
3. ***Multimedia mining***. There are many sites (Flickr, YouTube, etc.) for data exchange and media sharing. If there are tags or comments, analysis of multimedia content can be reduced to text analysis.
4. ***Text mining and data mining***. Social networks contain many textual information in various forms, for example, you can post comments, links to posts (messages), blogs or news articles. Sometimes users can mark each other, that is a form of textual information in the form of links, too. Tags (labels or keywords) that are describing different objects (images, text, or video) is of particular interest. Under this approach, the properties of tag streams, tagging models, tag semantics, visualizations and applications of tags are discussed. Many interesting issues are considered, such as the reason why people tag, what influences the choice of tags, kinds of tags, how to model the tagging process and how tags are created.

The **regulatory approach** involves the study of the trust level between participants, the norms, rules and sanctions that affect the behavior of participants and the processes of their interactions in the social network. In this case, the social roles, such as a relationship manager, and subordinate, friendship or family ties are explored. Because social networking is based on the interaction between various participants, it is natural to assume that this interaction affects the participants in terms of their behavior. The issues of this direction are as follows: how do we model the nature of influence across actors; how do we simulate the spread of influence; who is the most influential among the participants in the dissemination?

Social networks contain a large amount of personal information about the participants, for example, interests, information about friendship, demographic information, etc. This can lead to disclosure of different kinds of information in the social network. For this type of tasks, models on the basis of privacy mechanisms are useful.

Functional roles of participants are essential for effectiveness and sustainability of a social network, so the network can be used as a tool in order to identify experts in a particular field. The experts often form a network that matches the social network or organizational structure of the company.

Many complex problems require collective decisions by several experts. In such cases, we can more efficiently achieve a common goal when experts are cooperating with each other. There are many automated systems for experts' identification. For example, MITRE's Expert Finder is described in [3]. In addition to experts, brokers (leaders) are of interest in the social network analysis. Information brokers are people who play the role of mediators in a social network by linking a group of people, establishing communication between professionals, thus providing them with access to information.

A **dynamic approach** is a direction of social network analysis in which the objects of research are changes in the network structure over time: the network community is changing, new links emerge as new contacts are built and old links become obsolete as the members stop interacting, etc. This leads to changes in the structure of social networks in general and in individual communities. Important questions arise in this context: what are the laws which govern long term changes in a social network over time; are there any permanent configurations of a social network; how does a community evolve over time; what changes can occur and how we can track and represent them?

Link prediction in social networks is also an important problem. In most social networking applications, the links are dynamic and may change considerably over time. The prediction process may use either the network structure or the attribute-information at the different nodes. For such tasks, it is proposed to build a variety of structural and relational models [2].

Visualization provides a natural way to summarize the information in order to make it much easier to understand. It is important to create the algorithms that combine the methods of analysis and visualization techniques to improve understanding of the structure and dynamics of networks.

3. Some of the most famous social networks and software applications for social network analysis

Among major active social networking websites ([4], [5]) are Facebook, LinkedIn, Myspace, Twitter, Vkontakte, Odnoklassniki.ru, YouTube, etc.

Facebook. The network was founded in 2004 by Mark Zuckerberg. As of June 2012, Facebook had over 955 million active users [6], more than half of them using Facebook on their mobile devices. Facebook allows users to create a profile with a photo and information about themselves, invite friends, share messages, leave messages on their own and others' walls, upload photos and videos, create groups (community of interests), etc. It is possible to create applications for Facebook. (URL: `http://www.facebook.com`)

YouTube. Video hosting service was founded in 2005. Users can add, view and comment on videos, add annotations and captions to the video, etc. Due to its simplicity and usability, YouTube has become the most popular

video hosting community and the third site in the world in the number of visitors in June 2012 [7]. Sixty hours of video is uploaded every minute on YouTube. In January 2012, the number of daily video views on the site reached 4 billion. (URL: `http://www.youtube.com`)

LinkedIn. The social network was founded by Reid Hoffman in December 2002 and launched in May 2003. Basically the network is used for searching and establishing business contacts. According to data by February 2012, LinkedIn had more than 160 million users [8]. Slightly less than half of them are residents of the United States. (URL: `http://www.linkedin.com`)

Vkontakte. By March 2012, the audience was around 150 million [9], about 70% of them live in Russia. As is done in Facebook, the users of this network can send messages privately (via personal message) and publicly (through entries on "the wall", panels and meetings), track the activity of friends and communities via news feed. The network allows a user to share and download large files, as the technology used for distributed file sharing is BitTorrent, which makes Vkontakte one of the largest media archives of Runet. Odnoklassniki.ru, Facebook and other social networks use the Messaging Protocol XMPP (Extensible Messaging and Presence Protocol), formerly known as Jabber. (URL: `http://vk.com`)

Twitter was created in March 2006 by Jack Dorsey. The number of its registered users almost reached 500 million [10]. The system allows a registered user to send short text messages (up to 140 characters, known as "tweets") through the website interface, SMS, or a range of applications for mobile devices. The unregistered users can only read tweets. The distinctive feature of Twitter is public accessibility of posted messages, so this service is an online social networking and microblogging service. (URL: `https://twitter.com`)

Odnoklassniki.ru (Одноклассники in Russian, Classmates). The project started in 2006, its author is a Russian web-developer Albert Popkov. As of June 2011, the network registered more than 70 million users [11]. A feature of this network is that each user can see the names of all those who visited his/her profile, and all public actions of the users (forum posts, adding friends, photo upload etc.) are displayed in the available users' activity feed. Odnoklassniki.ru is the Russian counterpart of the American network Classmates.com. (URL: `http://www.odnoklassniki.ru`)

Flickr is an image and video hosting website and online community created in 2004. Flickr had a total of 51 million registered members and 80 million unique visitors [12] in June 2011. This network is one of the first Web 2.0 services. There is a possibility to add a title, a short description and keywords (tag) to each photo, for further search (URL: `https://www.flickr.com`).

There are many applications for modeling of interactions and processes in the network, for calculating the network parameters, and for graph vi-

sualization. For example, applications for visualization of the network Vkontakte (URL: http://www.yasiv.com/vk) or Facebook (URL: http://www.touchgraph.com/facebook) use different methods and algorithms described above.

The most popular tools of automatic analysis of social interactions include the following: NetMiner (URL: http://www.netminer.com/index.php), NetworkX (URL: http://networkx.lanl.gov), SNAP (URL: http://snap.stanford.edu), UCINet (URL: http://www.analytictech.com/ucinet), Pajek (URL: http://vlado.fmf.uni-lj.si/pub/networks/pajek), ORA (URL: http://www.casos.cs.cmu.edu/projects/ora), Cytoscape (URL: http://www.cytoscape.org), etc. An important requirement to such applications is the capability to handle very large amounts of data. As a result, processing is often parallelized.

There are applications that simulate the "theory of six handshakes", i. e. build a chain of connections (friends) between two users of a network: for the Russian network Vkontakte (URL: http://ienot.ru/hand) and for English networks (URL: http://www.sixdegrees.org, http://sixdegrees.com). The resulting chains are, indeed, quite short.

More information about existing applications for social network analysis can be found, for example, in [13], [14].

4. Models for social network analysis

One of the best known works in the social network theory is the theory on the spread of information in social networks known as "The Strength of Weak Ties" by an American sociologist Mark Granovetter. He has shown [15] that the weak links are much more effective than the strong ones for many social problems, such as job search. Weak ties are important sources of information, as they help to get more information about a participant or community. He called this effect "the strength of weak ties". The power of ties among participants is defined as a linear combination of duration, emotional intensity, intimacy or privacy, and the importance of mutual services that characterize the interaction and the corresponding edge of the graph.

Identifying experts in a subject and routing queries to them is an important problem, too. A model for expert identification and query routing in social networks based on the Ant Colony Optimization (ACO) approach is described in [16]. This is a probabilistic technique for finding good paths through graphs.

The original idea of the algorithm is based on the behavior of ants seeking a path between their colony and a source of food. Whenever an ant comes across a food source, it evaluates the quantity and the quality of the food source. The ant then goes back to the food source laying a trail of chemicals called pheromones in its path. The chemical trail gradually evaporates over

time. As other ants come across this trail, they follow this path to the food source and also deposit pheromones. Thus the frequently used trails become stronger as compared to the less frequently used. As the food source becomes depleted, the ants look for alternative routes and the old trails become weak as well.

In the proposed model, an ant is analogous to a query and the destination node is analogous to the expert. Thus a query trail keeps track of the routes taken by the queries of the network. The work begins with the placement of ants at the vertices, and they start moving. The direction is determined by the probabilistic method based on the formula:

$$P_i = \frac{l_i^q f_i^p}{\sum_{k=0}^{N} l_k^q f_k^p} \ ,$$

where P_i is the probability of moving on the way i,
l_i is the reciprocal of the weight (length) of the transition i,
f_i is the amount of pheromone deposited for transition,
q is the value which defines the "greed" of the algorithm,
p is the value which defines "gregariousness" of the algorithm $q + p = 1$.

The solution is not accurate and may be even one of the worst. However, due to the fact that the solution is stochastic, the repetition of the algorithm can give accurate results.

Another well known example of social network analysis is the experiment carried out by an American psychologist Milgram [17] in 1969. This experiment is called *Milgram's small world experiment* or *Six degrees of separation*. The hypothesis is that everyone is familiar with every other inhabitant of the planet through a chain of common acquaintances. This chain on average consists of six people. Today, this statement is not refuted. On the contrary, this hypothesis is supported by the fact that the diameter of most networks is relatively small. Let us consider the basic model of social network analysis.

Graph models of social networks are used for modeling the economic and communication links, analyzing the information propagation, community detection, etc.

Any social network can be mathematically represented as a graph $G = (V, E)$, where

V is the set of vertices;

E is the set of edges;

$|V| = N$ is the number of vertices in the graph.

The vertices of the social network graph are the participants, and the edges are relations between them. Relations can be directed and undirected. As a rule, we consider two basic types of relations: "friendship" (people are familiar with each other) and "interests" (there are common interests, people are in the same interest group).

There are three types of graph models [1]:

1. ***Stochastic block models*** are defined by a matrix A of size $N \times N$, where N is the number of groups of participants. Its element $a_{ij} \in [0,\ 1]$ shows the density of links between the participants belonging to a group v_i and the participants belonging to a group v_j. This graph contains no additional edges and vertices corresponding to relations of the participants within one group.
2. ***Probabilistic graph models*** are defined by a matrix A of size $N \times N$, where N is the number of participants. Its element $a_{ij} \in [0,\ 1]$ indicates the probability of interaction between a participant v_i and participant v_j for a certain period of time.
3. ***Regular graph models*** are defined by a matrix A of size $N \times N$.

Sometimes it is convenient to use the density coefficient for the analysis of graph models of social networks. It is defined as the ratio of the number of edges in a sample graph to the number of edges in the complete graph with the same number of vertices (complete graph is a graph in which all vertices are connected to each other). In addition, the network can be characterized by parameters such as the number of paths of a given length (a path is a sequence of vertices connected to each other), the minimum number of edges that divide the graph into several parts, etc.

Graph models of social networks are used for modeling the economic and communication links between people, analysis of information propagation, finding communities and associated sub-groups.

Analysis of centrality and other local network properties. There are various types of measures of centrality of a vertex within a graph that determine the relative importance of the vertex within the graph (i.e. how influential a person is within a social network). It should be noted that, of course, this is not a geometric visualization of centrality in the graph of relations. There are four measures of centrality widely used in network analysis [18]: degree centrality, betweenness centrality, closeness centrality, and eigenvector centrality.

Degree centrality is defined as the number of links incident upon a node (i.e., the number of ties that a node has):

$$C_D(v) = \deg(v).$$

Incoming links describe the popularity of a person, and outcoming ones show his sociability. The resulting value can be normalized by dividing by the total number of participants in the network.

In other words, the degree of centrality suggests that, among the more influential members of the network, there is one who has more friends, or one who belongs to a larger number of communities. However, the participant

having a large number of friends can be connected to the rest of the graph with a small number of edges. So, the following concept is defined.

Closeness centrality can be regarded as a measure of how fast the information is spread from one node to other nodes.

The distance between the two participants is the shortest path through the graph (geodetic distance). Thus, the direct participant's friends are at 1, friends of friends — at 2, friends of friends of friends — at a distance of 3, etc. The sum of distances is normalized. The resulting value is called the farness of a node v from the other nodes. The closeness is defined as the inverse of the farness.

$$C_C(v) = \frac{N-1}{\sum_{t \in V \setminus v} d_G(v,t)},$$

where $d_G(v,t)$ is the length of the shortest path from the node v to the node t.

In other words, the closeness centrality allows us to understand how close the participant is to all other participants in the network. Thus, not only the existence of direct friends is important, but also the presence of friends of friends is significant.

Betweenness centrality can be defined as a measure for quantifying the control of a human over communication between other humans in a social network. It is calculated as the number of the shortest paths between all pairs of vertices passing through that node.

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}},$$

where σ_{st} is the total number of the shortest paths from the node s to the node t;
$\sigma_{st}(v)$ is the number of those paths from the node s to the node t that pass through v.

The betweenness centrality can be normalized by dividing by the number of pairs of vertices not including v, which is $(N-1)(N-2)$ for directed graphs and $(N-1)(N-2)/2$ for undirected. Here N is the number of nodes in the giant component. The disadvantage of betweenness centrality is its computational complexity.

Eigenvector centrality is another measure of influence of a node in a network. Let the centrality of a participant be x_v, and the centrality of his direct friends (adjacent vertices) be x_j, x_k, x_l, etc. The eigenvector centrality is defined as the sum of centrality of adjacent vertices divided by a constant λ, i. e. $x_v = (x_j + x_k + x_l)/\lambda$. Writing the same equation for all friends, we get the vector of unknowns $X = (x_1, ..., x_v, ..., x_n)$. The addition rule is defined by the adjacency matrix $A = (a_{vt})$, i. e. $a_{vt} = 1$ if the vertex v is

linked to the vertex t, and $a_{vt} = 0$ otherwise. Next, we need to solve the equation $AX = \lambda X$, i.e. to find the eigenvalues and eigenvectors of the matrix A. This can be rewritten as follows:

$$C_E(v) = x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{vt} x_t,$$

where $M(v)$ is the set of neighbors of v;
λ is a constant.

Thus, if a participant has friends and their centrality is high, then his eigenvector centrality will be higher. The opposite is true: if the participant's centrality is high, then the centrality of his friends is higher. The downside to the eigenvector centrality is its computational complexity.

Katz centrality is a generalization of the degree centrality. The degree centrality measures the number of direct neighbors, and Katz centrality measures the number of all nodes that can be connected through a path.

$$C_{Katz}(i) = \sum_{k=1}^{\infty} \sum_{j=1}^{N} \alpha^k (a^k)_{ji},$$

where $\alpha \in (0,\ 1)$ is a contribution of a distant node (attenuation factor).

Katz centrality can be viewed as a variant of the eigenvector centrality:

$$C_{Katz}(i) = x_i = \alpha \sum_{j=1}^{N} a_{ij} (x_j + 1).$$

Centrality can be computed applying the reference ranking algorithm (***PageRank algorithm***), used by the Google Internet search engine. The rank value indicates importance of a particular webpage. A hyperlink to a page is counted as a vote of support. The PageRank of a page is defined recursively and depends on the number of all pages linked to it ("incoming links") and their PageRank metric. A page that is linked to many pages with a high PageRank receives a high rank itself.

$$C_{PageRank}(i) = x_i = \alpha \sum_{j=1}^{N} a_{ji} \frac{x_j}{L(j)} + \frac{1-\alpha}{N},$$

where $L(j) = \sum_j a_{ji}$ is the number of neighbors of the node j (or the number of outbound links in a directed graph).

Compared to the eigenvector centrality and Katz centrality, one major difference is the scaling factor $L(j)$. Another difference between PageRank and the eigenvector centrality is that the PageRank vector is a left hand

eigenvector (note that the factor a_{ji} has indices reversed). The predecessor of the PageRank algorithm is ***Hyperlink-Induced Topic Search (HITS)*** algorithm proposed by Kleinberg [19]. There are many other centrality indices.

A useful technique of social network analysis is the ***TrustRank algorithm*** described in [20]. TrustRank is a method for determining relative importance of web documents. Initially, this algorithm was designed for separating useful webpages from spam. Experts manually evaluate the TrustRank of a small number of sites for the reference sample which can be considered as reliable. The algorithm is based on the assertion that useful sites rarely link to spam, but the pages that contain spam often have links to useful websites. The more links to the site, the less trust is passed to each link. The credibility of the website decreases with increasing the distance between it and the reference sample.

The structural position power and expert prestige power are key indicators in determining the differences in resources of the network members. ***Genuine Progress Indicator (GPI)*** was defined for measuring this characteristic in the network exchange theory [21]. The network exchange theory assumes that GPI can predict power and profit rankings in exchange networks by detecting structural advantage or disadvantage of a position.

$$GPI_i = \sum_{k=1}^{g-1} (-1)^{k-1} P[\,i\,]_k,$$

where $P[\,i\,]_k$ is the number of non-intersecting paths of length k, passing through the node v_i. The participant's strength of v_i compared to the participant's strength of v_j is calculated as $GPI_{ij} = GPI_i - GPI_j$.

The methods of community detection and analysis of connected subgroups. The connected subgroups (community) in a network are characterized by a large number of links among their participants and much smaller number of connections with other participants. Community detection allows us to study the stability of social structures. The simplest case of a connected group is a clique – community where all members are associated with each other, and other members of the network can not be included in this group, because they do not have relations with all members of the community. Thus, a clique is a maximal complete subgraph. If we analyze dissemination of information, we can give another definition of community. Community is a set of participants, where a path between any two of them does not contain more than one intermediate vertex. As a result, information is transferred from one participant to another with minimal distortion in a connected group. Connected groups can also be separated by multidimensional scaling or factor analysis of the connection matrix [1].

The following technique is used for analysis of stability of a group structure over time. First, a three-dimensional matrix is constructed, the rows of which are estimates (which are given by participants) of interactions with all other participants, the columns are the participant's own estimates of interactions, the third axis is time periods. Next, a graph showing a change in the structure of subgroups over time can be built.

After that, the methods of reducing the dimension of data (e. g., the principle of main components) are used, i. e. we consider the projection of vertices of a low-dimensional Euclidean space for description of relationships between rows and columns of the matrix. As a result, changes in the user's status in the network can be visualized on the background of changes in the status of subgroups [22].

The resulting projection can be clustered using the standard clustering algorithms: hierarchical and statistical. The advantage of the hierarchical methods is the possibility of presenting the result of clustering in a dendrogram. The main difficulty of such methods is to find the right measure of distance (the shortest path between nodes) or measure of similarity of two vectors. The most commonly used measures of similarity are cosine similarity (also known as Ochiai coefficient) and Jaccard coefficient. There exist top-down clustering techniques and bottom-up clustering techniques.

A more detailed review of algorithms for community detection can be found, for example, in [23].

Structural equivalence of the participants. This approach is contrary to the study of connected groups. Participants are equivalent when they occupy the same position in the social structure of the network, i.e. the structure and the type of interactions of these participants with others are equivalent, and equivalent participants should not interact with each other. As a measure of equivalence, we may consider density of ties with structural subgroups of the network participants [24]. Along with structural equivalence, regular equivalence of participants is used to identify nodes that are playing the same structural role, even if they are not connected to each other.

In order to determine the structural equivalence of two participants, it is necessary to compare the structure of their interactions with other members, that is, to compare the corresponding columns of the relationship matrix. This can be done by calculating the distance between these vectors (e. g., Euclidean or Chebyshev metric) or coupling coefficients (e. g., Pearson correlation). Incoming and outgoing edges should be considered for directed graphs, so the two corresponding matrices are considered simultaneously [1].

Role algebras. This method of social network analysis is aimed at revealing the logic of interactions of participants in a block model. This makes it possible to identify the same principles of relations between participants in social networks.

Let us define, for example, a matrix of likes and dislikes as follows:

$$LIKE = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, DISLIKE = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Now we can analyze the combinations of interactions of the participants by multiplying the corresponding matrices.

Analysis of dyads and triads. A dyad is a set of two network members (vertices) and all interactions (edges) between them. Dyads can be in one of four states for each type of interaction: there is no connection between the participants, a link is directed from the first participant to the second, a link is directed from the second participant to the first, there are interconnections between the participants. Dyads analysis helps us to determine the probability of having an edge between them, the degree of dependence of participants on their properties, to define the conditions and the direction of information transmission, etc.

In addition to the above, the issue of interactions transitivity is investigated for the triads (there are three interacting participants). A relation is transitive if, whenever it relates some v_1 to some v_2, and that v_2 to some v_3, it also relates that v_1 to that v_3. Balance is also an important parameter that describes the local links of the participants, and often used in the analysis of dyads and triads. Balance is the lack of situations such as "positive interaction (friendship, partnership) between v_1 and v_2, between v_1 and v_3, but negative interaction (enmity, rivalry) between v_2 and v_3". There is an assumption in [25] that a balanced network is psychologically more comfortable for participants and more stable as compared to unbalanced. These characteristics describe local connections of the participants and they are often used in the analysis of dyads and triads.

Models of network dynamics. Approaches described in [22], [26], are used to study the network dynamics.

Modeling of graph evolution is a direction which studies a wide variety of network formation strategies and shows that edge locality plays a critical role in the evolution of networks. For example, in [27] it was observed that networks become denser over time. At the same time, the number of edges grows superlinearly with the number of vertices. Moreover, densification follows a power-law pattern. Another paper [28] presents algorithmic tools for the analysis of evolving networks and focuses on assessing the community affiliation of users and how this assessment changes over time. The proposed algorithms are based on dynamic programming, exhaustive search, maximum matching, and greedy heuristics. There is another surprising observation in this paper: the network diameter often decreases over time, in contrast to a current opinion that such distance measures should increase slowly as a function of the number of vertices.

The focus is on determining the approximate cluster of users and their temporal changes. The principle of Minimum Description Length (MDL)

is used for finding patterns in data and detecting communities in dynamic networks.

Dynamic graph mining is sometimes convenient to carry out on the basis of the paradigm of association rule extraction and analysis of frequency models. A novel type of frequency-based patterns and graph evolution rules is defined. Then the problem of searching for typical patterns of structural changes in dynamic networks is considered. First, calculate a set of frequent graph patterns that describes specific evolutionary mechanisms, and then find the graph evolution rules that satisfy a given minimum confidence constraint. Also, the methods of identifying the subgraphs changing over time by means of vertex importance scores and vertex-closeness changes are used for analysis of the graph of a network. The most relevant subgraph is not the most frequent, but the most significant one. The history of an edge in a dynamic graph can be represented as a series of zeros and ones corresponding to the presence or absence of the edge. Then, to obtain the frequency graph, the model uses traditional methods to describe the graph as a sequence of ones and zeros. Then conventional graph mining techniques are applied to mine the frequent patterns.

The problem of ***link formation prediction*** considers whether two particular nodes are likely to become connected in the future. The parameters based on the link structure of the network are considered for solving this problem, such as the number of common neighbors, geodesic distance, and hitting time in the social network. There are several models relying on machine learning techniques and using personal information of users (music, books, etc.) to improve the accuracy of predictions. Several probabilistic models, such as Markov logic, relational Markov networks, Markov random fields, and probabilistic relational models have been used to capture the relations existing in data. Other approaches are based on properties of the users themselves. According to [29], many connections in large social networks (the blogosphere, in this case) can be explained by matching demographic groups, topical interests in common, or geographical proximity.

Methods based on ontologies. It is possible to estimate the parameters of social networks (diameter, number of participants, average path length, etc.) by means of ontologies [30]. First, analysis of kinds of network elements is performed: people, objects (music, photos, video, messages, etc.) and interactions (knows, reports, comments, etc.).

Then the FOAF (Friend of a friend) ontology is applied to identify the participants of a social network and the content added by them. This ontology describes people, their activity and relations with other people and objects. The description of social relationships between people in the FOAF is based on the transitivity of trust. FOAF ontology specification is available at `http://xmlns.com/foaf/spec`.

SCOT (Social Semantic Cloud of Tags) ontology is used to describe tags,

for instance, in [30], [31]. The SCOT ontology provides a model for expressing the main concepts and properties required to describe information for tagging activities (e. g., users, tags, resources, etc.) in the Semantic Web. The SCOT ontology specification is available at `http://rdfs.org/scot/spec`.

Two ontologies have been created: SemSNA (Semantic Social Network Analysis) and SemSNI (Semantic Social Network Interactions). SemSNA describes the concepts of social network analysis, while SenSNI is used for representation of interactions in a social network. The description of SemSNA and SemSNI ontologies is presented in [30]. For example, SemSNA can be used for computing centrality of nodes; SemSNI can be used for representation of: page visits, comments, private messages, etc. As part of semantic network analysis, it became possible to calculate the parameters of subgraphs by means of semantic links ("Family", "I like", "Friend") and types of interactions ("Comment", "New message", etc.).

5. Conclusion

There exist different approaches to the analysis of social computer networks, which gives rise to the problem of combining the results of the research. Therefore, the actual problems of network analysis include the following: creation of a unified theory of social networks, creation of a generic set of measures of distances to determine the distance between elements in a network, and systematization of different measures of completeness of a network.

New methods of statistical analysis and their combination with the graph theory algorithms can be useful in the study of the attributes of users, the links between participants, and pattern detection in networks. Sometimes the relationships between participants can be regarded as the stochastic characteristics that describe the evolution of networks. The problem of finding a person in a social network can be compared, in a sense, to the problem of finding relevant documents in a collection of documents with references. Thus, many text processing techniques can be adapted for social network analysis.

Less common methods include, for example, the use of topology tools. In particular, a study of social networks proposed in [1] uses Koenig's theorem, which states that any graph can be laid without self-intersections on a compact orientable topological surface. This gives us the opportunity to consider the geometry of the social space from new positions.

It has been suggested in [2] to determine the analytical structure of a network in which the participants are adversaries, and the relationships among different adversaries may not be fully known. This type of networks is much more difficult to study because connections can not be established a priori. It would be interesting to explore this kind of relations for analytical purposes.

In order to generalize the behavior of network participants to the entire

network, it is necessary to study the methods of detection and characterization of networks, distribution patterns of these characteristics, and creation of techniques for identifying the causes of interactions via the structure of the social network. These processes are particularly important in the analysis of modern social networks with many participants.

References

[1] Churakov A.N. Social network analysis. // Sociological studies. – 2001. – N 1. – P. 109–121 (in Russian).

[2] Charu C. Social network data analytics. – Springer, 2011. – 520 p.

[3] Maybury M., D'Amore R., House D. Awareness of organizational expertise. // International Journal of Human-Computer Interaction. – 2002. – Vol. 14, N 2. – P. 199–217.

[4] http://ru.wikipedia.org/wiki/Социальная_сеть (In Russian).

[5] List of social networking websites. – http://en.wikipedia.org/wiki/List_of_social_networking_websites

[6] Sengupta S. Facebook's prospects may rest on trove of data. // The New York Times. – 14.05.2012. – http://www.nytimes.com/2012/05/15/technology/facebook-needs-to-turn-data-trove-into-investor-gold.html?_r=0

[7] Alexa – The Web Information Company. – 2012. – http://www.alexa.com/siteinfo/youtube.com

[8] LinkedIn – Press Center. – http://press.linkedin.com/about

[9] Vkontakte. (in Russian) – http://vk.com/catalog.php

[10] Twitter has 500 million registered users. – http://www.mediabistro.com/alltwitter/500-million-registered-users_b18842

[11] Blagoveschensky A. Number of users of "Classmates" has exceeded 70 million. // The Russian newspaper. – 2011. – 21 June. (In Russian) – http://www.rg.ru/2011/06/21/odnoklassniki-site-anons.html

[12] Yahoo Yahoo! Advertising solutions. – http://advertising.yahoo.com/article/flickr.html

[13] Prokhorov A., Larichev N. Computer visualization of social networks. // ComputerPress. – 2006. – N 9. – P. 156–160 (In Russian).

[14] http://en.wikipedia.org/wiki/Social_network_analysis_software

[15] Granovetter M.S. The Strength of Weak Ties. // American Journal of Sociology. – 1973. – Vol. 78, N 6. – P. 1360–1380.

[16] Ahmad M., Srivastava J. An Ant Colony Optimization Approach to Expert Identification in Social Networks. // Proc. of the First International Workshop on Social Computing, Behavioral Modeling, and Prediction. – 2008. – P. 120–128.

[17] Milgram S. The Small World Problem. // Psychology Today. – 1967. – Vol. 2. – P. 60–67.

[18] `http://en.wikipedia.org/wiki/Centrality`

[19] Kleinberg J. M. Authoritative sources in a hyperlinked environment. // Journal of the ACM. – 1999. – Vol. 46, N 5. – P. 604–632.

[20] Gyöngyi Z., Garcia-Molina H., Pedersen J. Combating Web Spam with TrustRank. // Proc. of the XXX Internat. Conf. on Very Large Data Bases. – 2004. – P. 576–587.

[21] Davern M. Social networks and economic sociology: A proposed research agenda for a more complete social science. // American Journal of Economics & Sociology. – 1997. – Vol. 56, Iss. 3. – P. 287–302.

[22] Bonchi F., Castillo C., Gionis A., Jaimes A. Social Network Analysis and Mining for Business Applications. // ACM TIST. – 2011. – Vol. 2, Iss. 3. – P. 22–58.

[23] Fortunato S. Community detection in graphs. // Physics Reports. – 2010. – Vol. 486, Iss. 3–5. – P. 75–174.

[24] Wasserman S., Faust K. Social Network Analysis: Methods And Applications. – Cambridge University Press. – 1994. – 825 p.

[25] Johnson J., Ironsmith M. Assessing children's sociometric status: Issues and the application of social network analysis. // Journal of Group Psychotherapy, Psychodrama & Sociometry. – 1994. – Vol. 47, Iss. 1. – P. 36–49.

[26] Hanneman R. Computer-Assisted Theory Building: Modeling Dynamic Social Systems. – Riverside, University of California. – 1988. – Published in digital format. – `http://faculty.ucr.edu/~hanneman`

[27] Leskovec J., Kleinberg J., Faloutsos C. Graphs over time: Densification laws, shrinking diameters and possible explanations. // Proc. of the XI ACM SIGKDD Internat. Conf. on Knowledge Discovery in Data Mining. – New York, 2005. – P. 177–187.

[28] Tantipathananandh C., Berger-Wolf T., Kempe D. A framework for community identification in dynamic social networks. // Proc. of the XIII ACM SIGKDD Internat. Conf. on Knowledge Discovery and Data Mining. – New York, 2007. – P. 717–726.

[29] Kumar R., Novak J., Raghavan P., Tomkins A. Structure and evolution of blogspace. // Communications of the ACM. – 2004. – Vol. 47, N 12. – P. 35–39.

[30] Erétéo G., Gandon F., Buffa M., Corby O. Analysis of a Real Online Social Network Using Semantic Web Frameworks. // Proc. of the VIII Internat. Semantic Web Conf. – 2009. – P. 180–195.

[31] Kim H. L., Yang S.-K., Song S.-J., Breslin J. G., Kim H.-G. Tag Mediated Society with SCOT Ontology. // Proc. of the VI Internat. Semantic Web Conf. – 2007. – P. 295–302.